УДК 811.93

О. В. КАНІЩЕВА, М. В. ЧУХНЕНКО

АВТОМАТИЧНИЙ ПОШУК КЛЮЧОВИХ СЛІВ У КОРПУСІ МАСОВОЇ ЛІТЕРАТУРИ

Проаналізовані основні методи пошуку ключових слів у лінгвістичних корпусах, описані сфери їх застосування та їхні властивості. Розглянуто основні переваги і недоліки методів – кількісний метод аналізу, якісний метод аналізу та статистичний метод, які використовуються у сучасній корпусній лінгвістиці. Розроблено корпус масової літератури та описана його структура. Також було розроблено програмне забезпечення, яке реалізує автоматичний пошук ключових слів у створеному корпусі. Проведено аналіз отриманих результатів роботи програми.

Ключові слова: корпус, корпусна лінгвістика, масова література, статистичний аналіз, літературні жанри, ключові слова, Weirdness.

О. В. КАНИЩЕВА, М. В. ЧУХНЕНКО

АВТОМАТИЧЕСКИЙ ПОИСК КЛЮЧЕВЫХ СЛОВ В КОРПУСЕ МАССОВОЙ ЛИТЕРАТУРЫ

Проанализированы основные методы поиска ключевых слов в лингвистических корпусах, описаны сферы их применения и их свойства. Рассмотрены основные преимущества и недостатки методов – количественный метод анализа, качественный метод анализа и статистический метод, которые используются в современной корпусной лингвистике. Разработан корпус массовой литературы и описана его структура. Также было разработано программное обеспечение, которое реализует автоматический поиск ключевых слов в созданном корпусе. Проведено анализ полученных результатов работы программы.

Ключевые слова: корпус, корпусная лингвистика, массовая литература, статистический анализ, литературные жанры, ключевые слова, Weirdness.

O. V. KANISHCHEVA, M. V. CHUKHNENKO

AUTOMATIC SEARCH OF KEYWORDS IN THE BELLES-LETTRES CORPUS

The basic methods of search of key words in linguistic corpus have been analyzed, the spheres of their application and their properties have been described. The main advantages and disadvantages of the methods used in modern corpus linguistics – quantitative analysis method, qualitative analysis method and statistical method have been considered. The corpus of mass literature has been developed and its structure has been described. Besides formula of Weirdness which has been used in the software has been described. Also, software that implements the automatic search of keywords in the created corpus has been developed. The results of the program have been analyzed.

Keywords: corpus, corpus linguistics, mass literature, statistical analysis, literary genres, keywords, Weirdness.

Вступ. Розвиток корпусної лінгвістики, а також побудова корпусів є однією з актуальних проблем сучасного мовознавства. Корпус в лінгвістиці – це сукупність текстів, яка зібрана в єдине ціле за

певними, відповідним до конкретної дослідницької задачі, критеріям і відображає ту чи іншу сферу використання мови. У сучасній лінгвістиці під корпусом зазвичай мають на увазі корпус текстів в електронній формі.

На даний момент використання корпусів відіграє провідну роль при проведенні більшості лінгвістичних досліджень. Традиційні способи збору мовних даних включали перегляд і ручну обробку письмових текстів, яка полягає в виписці з них потрібних прикладів, опитування інформантів з подальшим вивченням польових анкет, запис текстів в письмовій формі, створення словникових картотек. Ця попередня діяльність, а також процес поточної обробки картотеки, пошуку потрібних одиниць були дуже трудомісткими, забирали чимало часу і не дозволяли обробляти великі масиви матеріалу. Крім того, при традиційній технології збору та обробки мовних даних існувала проблема ергономічного поновлення зібраного матеріалу, а також була відсутня можливість доступу до мовних даних на відстані.

Нові інформаційні технології та технічні засоби значно полегшили збір мовних даних. Тепер обмежень на обсяг аналізованого матеріалу і швидкість пошуку інформації в ньому немає, а це означає, що в розпорядженні дослідника виявляються колосальні масиви текстів самого різного типу. Це не забарилося позначитися на розвитку знань про мову: масова – в тому числі статистична – обробка текстів, недоступна раніше, дозволила виявити в структурі і розвитку мови такі закономірності, про існування яких наука раніше лише здогадувалася, але не могла їх обґрунтувати.

На корпусі тестуються системи автоматичної обробки тексту і перевіряються різні лінгвістичні теорії. Одним із завдань комп'ютерної корпусної лінгвістики є завдання автоматичного визначення ключових слів тексту.

Аналіз останніх досліджень і публікацій. Наявність великої кількості текстів в електронній формі суттєво полегшило завдання створення великих представницьких корпусів розміром в десятки і сотні мільйонів слів, але не ліквідувало проблем: збір тисяч текстів, зняття проблем з авторськими правами, приведення всіх текстів в єдину форму, балансування корпусу за темами та жанрами забирають багато часу [1].

Авторами статті були проаналізовані сучасні існуючі лінгвістичні корпуси англійської мови, такі як: American National Corpus [2], Bank of English [3], British National Corpus [4], Corpus of Contemporary American English [5], International Corpus of English [6], Oxford English Corpus [7] та Scottish Corpus of Text and Speech [8]. Розглянуті корпуса в цілому поєднують в собі декілька сфер мови: художня література, газети, журнали, усне мовлення, тощо. Не існує такого корпусу, який б містив тільки художню літературу.

У сучасних корпусних лінгвістиці використовуються два основні методи: кількісний метод аналізу та якісний метод аналізу.

Найважливіша відмінність між цими двома методами полягає не в тому, які проблеми досліджуються, а в тому, яким чином вони досліджуються. Наприклад, вплив англіцизмів та галіцізмів на німецьку мову досліджували два учасника – Шанке (2001) і О'Халлоран (2002). Обидва працювали із власними корпусами, складеними із текстів газет і журналів. Корпус Шанке містить всі випуски німецької газети Handelsblatt за березень 2000 р. а корпус О'Халлоран включає корпус мови моди (випуски німецького журналу для жінок Brigitte, датовані різними роками) і корпус стандартної мови (що відносяться до різних років випуску німецького журналу Stern і німецької газети Berliner Illustrierte Zeitung). Шанке застосовує якісний метод аналізу, О'Халлоран – кількісний.

Мета при використанні якісного методу полягає в тому, щоб виявлені іноземні слова класифікувати за частинами мови і віднести їх до різних тематичних галузях, наприклад, комп'ютер, біржа, банківська справа. Якісний аналіз передбачає виявлення, класифікацію та інтерпретацію феноменів. При кількісному аналізі вивчалося поширення англіцизмів та галіцізмов протягом останніх 100 років. У результаті аналізу було встановлено, що кількість типів іноземних слів у корпусі зростає, а саме: з 0,6% в рік (1902) до 2,0% в рік (1997). Крім того, частка іноземних слів (точніше, їх словоформ) в корпусі мови (14%) в будь-який момент часу перевищує частку іноземних слів в корпусі стандартного мови (4%).

Мета застосування кількісного методу полягає в тому, щоб виявити частотність певних феноменів і порівняти їх з метою підведення підсумків дослідження. За кількісними показниками стандартним чином обчислюється величина корпусу, вона вимірюється в текстових словах і є найважливішою вихідною (основною або умовною) величиною для кількісного аналізу. Якщо величина корпусу невідома, то кількісний аналіз має сенс тільки в тому випадку, коли можна порівняти результати для багатьох подібних феноменів всередині корпусу [9].

Також ми дослідили статистичні методи, які використовує корпусна лінгвістика. Статистичні методики на базі корпусу використовуються для розробки, настройки та тестування різних систем, заснованих на використанні комп'ютерних технологій (машинний переклад, розпізнавання і синтез мови, інформаційний пошук, засоби перевірки орфографії і граматики, тощо). Корпусна лінгвістика зробила можливим уточнити результати і висновки, проведених раніше досліджень мовлення та провести нові, більш широкі і системні за охопленням емпіричного мовного матеріалу лінгвістичні дослідження [10].

Застосовуючи статистичні методи дослідження із використанням корпусів можна або підтвердити, або спростувати припущення про мовні явища.

Статистичними методами на матеріалі корпуса можна визначити які слова регулярно зустрічаються разом і таким чином їх можуть віднести до стійких словосполучень. Стійкі словосполучення являє собою із семантичної точки зору неподільну смислову одиницю, що дуже важливо ураховувати

у лексикографії та системах автоматичної обробки тексту. Корпуси – це багаті джерелом даних для досліджень з лексикографії та граматики. З дослідженням по лексикографії тісно пов'язані дослідження в галузі семантики. Спостерігаючи оточення тієї або іншої лінгвістичної одиниці у корпусі, можна встановити певні семантичні ознаки, які характеризують дану одиницю [9].

У лінгвістиці широко використовуються методи описової статистики. Описова статистика – один із розділів статистичної науки, в рамках якої вивчаються методи опису і представлення основних властивостей даних. Дозволяє узагальнювати первинні результати, отримані при спостереженні або в експерименті. Застосування описової статистики включає наступні етапи: збір даних, категоризація даних, узагальнення даних, подання даних. Протиставляється статистичному висновку в тому сенсі, що не чинить висновків про генеральну сукупність на підставі результатів дослідження окремих випадків. Статистичний висновок же передбачає, що властивості і закономірності, виявлені при дослідженні об'єктів вибірки, також притаманні генеральної сукупності [11].

Метою дослідження є вирішення завдання автоматичного виявлення ключових слів у жанрових корпусах на основі створеного корпусу англійської мови, для їхнього подальшого використання при вирішені задачі автоматичного визначення жанру тексту.

Матеріали і результати дослідження. Для вирішення поставленої задачі було розроблено корпус масової літератури Belles-lettres письмової англійської мови, який включає в себе такі чотири жанри масової літератури: пригодницький роман (Adventure), детектив (Detective), фентезі (Fantasy) та наукова фантастика (Science-fiction).

Масова література – багатомовний термін, що має декілька синонімів: популярна, тривіальна, бульварна література; традиційно цим терміном позначають ціннісний «низ» літературної ієрархії – твори, що відносяться до маргінальної сфери загальновизнаною літератури, які заперечують як кітч, псевдолітература. Нерідко під масовою літературою розуміють весь масив художніх творів певного культурно-історичного періоду (або будь-якого літературного напряму), які розглядаються як фон вершинних досягнень письменників першого ряду [12].

Для кожного жанру масової культури створюється певний звід законів та правил – модель, яка забезпечує їх впізнання, тому читач ніколи не відчуває розчарування від нездійснених очікувань: у детективі злочинця завжди буде викрито, героїня жіночого роману знайде своє щастя, тощо. На відміну від елітарної культури масова культура говорить зі своїм читачем на зрозумілій йому мові.

Авторами статті було створено корпус за обсягом – 135147 слів. Розмір текстових файлів приблизно 5500 слів. Існує чотири папки, що відповідають відповідним жанрам. В них тексти зберігаються у текстових файлах із розширенням .txt у кодуванні UTF-8.

Для кожного жанру масової літератури було вибрано найяскравіших представників цих жанрів. Тексти обирались по декілька глав або одна глава із кожного твору, але при цьому намагались обрати ті глави, які містять найбільшу кількість потрібної для дослідження лексики.

Для пригодницького роману обрали наступних авторів з їх творами: Р.Л. Стівенсон «Острів скарбів» (англ. Treasure Island), Р. Сабатіні «Одисея капітана Блада» (англ. Captain Blood), Ж. Верн «П'ять тижнів на повітряній кулі» (англ. Five Weeks in a Balloon), Дж. Лондон «Серця трьох» (англ. Hearts of Three), Ж. Верн «Таємничий острів» (англ. The Mysterious Island) та Ж. Верн «Навколо світу за вісімдесят днів» (англ. Around the world in eighty days). Усі вони знаходяться у файлах із назвами A1 – A2 відповідно.

Для детективу обрали А.К. Дойл «Берилова діадема» (англ. The Adventure of the Beryl Coronet), Е.А. По «Вбивства на вулиці Морг» (англ. The Murders In The Rue Morgue), А. Крісті «Загадкова пригода в Стайлзі» (англ. The Mysterious Affair At Styles), А. Крісті «Десять негренят» (англ. And Then There Were None), А.К. Дойл «Собака Баскервілів» (англ. The hound of the Baskervilles), А.К. Дойл «Етюд в багряних тонах» (англ. A Study in Scarlet). Усі вони знаходяться у файлах із назвами D1 – D2 відповідно.

Для фентезі обрали Дж.К. Роулінг «Гарі Потер та Орден Фенікс» (англ. Harry Potter and the Order of the Phoenix), Дж. Толкін «Володар Перснів» (англ. The Lord of the Rings), К.С. Льюїс «Хроніки Нарнії: Лев, Біла Відьма та шафа» (англ. The Chronicles of Narnia: The Lion, the Witch and the Wardrobe), Ф. Пулман «Чудовий ніж» (англ. The Subtle Knife), Н. Гайман «Зоряний пил» (англ. The Stardust), Р. Дал «Відьми» (англ. The Witches). Усі вони знаходяться у файлах із назвами F1 – F2 відповідно.

Для наукової фантастики обрали І. Єфремов «Туманність Андромеди» (англ. Andromeda A space-age tale), О. Толстой «Аеліта» (англ. Aelita), Л. Станіслав «Соляріс» (англ. Solaris), Г.Дж. Уелс «Війна світів» (англ. The War of the Worlds), Ж. Верн «Навколо Місяця» (англ. All Around The Moon), Е. Гамильтон «Зоряні королі» (англ. The Star Kings). Усі вони знаходяться у файлах із назвами SF1 – SF2 відповідно.

Розроблений корпус використовується для пошуку ключових слів. Ключові слова – це слова, які вживаються незвичайно часто в порівнянні з будь-яким референтним корпусом. Для їх визначення кожен підкорпус порівнюється із трьома іншими.

Крім загальнонаукового розуміння ключових слів, що визначають зміст тексту і передають його основний зміст, даний феномен розглядається такими науковими і прикладними дисциплінами як психолінгвістика, теорія комунікації, комп'ютерна та когнітивна лінгвістика, інформатика.

У результаті систематизації даних різних дослідників виділено перелік істотних властивостей і функцій ключових слів у текстах, які є значущими в контексті моделювання та алгоритмізації процесу їх вилучення.

Отже, ключові слова характеризуються тим, що:

- є найбільш вживаними (частотними) найменуваннями, позначають ознаку предмета, стан або дію;
- представлені значущою лексикою, досить узагальнені за своєю семантикою (середнього ступеня абстракції), стилістично нейтральні;
- пов'язані один з одним мережею семантичних зв'язків, перетину значень;
- більше половини слів ядра тематичного компонента складається з ключових слів, а мінімальний набір ключових слів наближається до інваріанта змісту при їх логічному впорядкування;
- набір ключових слів складається з 5-15 або 8-10 слів, що відповідає об'єму оперативної пам'яті людини, в тексті міститься 25-30% ключових слів;
- набір ключових слів визначає контексти слів, що володіють максимальною передбачуваністю.

У процесі сприйняття тексту ключові слова виділяють за синтаксичною позицією (заголовок або перше речення), по частотності вживання, лексичним патернам, незвичайним сполученням, відносинам синонімії, антонімії, морфологічної та семантичної похідності.

Так як наведені характеристики ключових слів проявляються на декількох рівнях розгляду тексту – морфологічному, лексичному, синтаксичному та прагматичному, то їх розпізнавання має на увазі відносну складність використовуваних методів і багатоетапність, що реалізує їх алгоритми. Основні методи вирішення даного завдання, будучи статистичними, базуються на обчисленні різних частотних характеристик тексту [13].

Щоб знайти ключові слова у корпусі, зазвичай порівнюють два корпуси між собою: один корпус – спеціалізований, тобто він містить певну лексику, яка інтересує дослідників, в ньому як раз потрібно знайти ключові слова, другий корпус – референтний, тобто він містить загальну лексику. Як правило референтний корпус має бути більшим за розміром, ніж спеціалізований.

Якщо слово зустрічається наприклад 5% у спеціалізованому корпусі, а у референтному 6%, то таке слово не буде ключовим. Але якщо слово відповідно зустрічається 25% та 6%, то таке слово буде ключовим.

Поруч з кожним ключовим словом знаходяться різні цифри, які містять інформацію про те, як часто вживається кожне слово в вихідному тексті (текстах) і наскільки ця частотність відрізняється від частотності його вживання в референтному корпусі.

Відмінності в розподілі певних лексичних одиниць та їх варіантів в спеціалізованих та загальних (референтних) корпусах можна кількісно визначити за відносними частотами спеціалізованого корпусу та загального (референтного) корпусу. Таке співвідношення називають індексом Weirdness («дивності») спеціалізованого корпусу. Таку Weirdness («дивність») використовують акцентований, і, можливо, ексцентричний вибір лексичних предметів, виміряних з точки зору їх частоти виникнення.

Найбільш «дивні» слова в тексті схильні представляти його більш точно, ніж ті, які не такі «дивні». Якщо співвідношення рівне, то лексична одиниця має однакову частоту як у загальному (реферативному), так і спеціалізованому корпусі; якщо співвідношення є більшим, то цей елемент використовується частіше в спеціалізованому корпусі.

Можна стверджувати, що зіставлення частотного розподілу лексичних одиниць в спеціалізованому та загальному корпусах може виявити ключові слова для спеціалізованого корпусу. Ця методика має перевагу в тому, що слова «закритого класу», як правило, мають співвідношення приблизно 1:1 у цьому порівнянні, тоді як спеціалізований термін – ключові слова, а не усі слова із корпусу, – матимуть набагато більший коефіцієнт, оскільки їх частота в загальному корпусі буде низькою або потенційно нульовою. Формула для розрахунку індексу weirdness представлена нижче:

$$Weirdness = \frac{w_s / t_s}{w_g / t_g}, \qquad (1)$$

де w_s – частота слів у спеціалізованому корпусі, t_s – загальна кількість слів у спеціалізованому корпусі, w_g – частота слів у загальному (референтному) корпусі, t_g – загальна кількість слів у загальному (референтному) корпусі [14].

Для визначення ключових слів для наших підкорпусів було розроблено наступний алгоритм:

1) Видалення пунктуації з текстів.

2) Для кожного слова жанрового підкорпусу (спеціалізований корпус) рахується відносна частота.

3) Рахується відносна частота для кожного слова із загального корпуса (референтний).

4) Розраховується індекс Weirdness.

Для реалізації програмного забезпечення пошуку ключових слів у створеному корпусі англійської мови була обрана високорівнева мова програмування Python.

Вивід результатів обмежили для того, щоб виводилось лише 75% слів. Це було зроблено для того, щоб вивести якомога більше слів, які на нашу думку, є ключовими. Спочатку визначили максимальне значення із результату, який отримали за рахунок формули Weirdness, та визначили 75% для кожного результату.

У результаті ми отримали 4 файли у форматі .xlsx. Найкращий результат досягнуто у жанрі наукової фантастики – найбільша частина слів є характерними для цього жанру: *space, cosmos, Earth, cosmic, lunar, solar, planetary, planet*, тощо. На другому місці – пригодницький роман із словами: *treasure, isle, anchor, sail, buccaneers, sailor, frigate*, тощо. На третьому місці – детектив із словами: *police, inspector, cooperation, thief, theft, confession, guilt, custody,*

confess, тощо. Найгірший результат – фентезі зі словами *dwarf, elf, wizard, hobbit, charm*, тощо.

В результатах є багато «шуму», тобто слів, які не характеризують певний жанр, а які відносяться до загальновживаної лексики. Такі результати зумовлені тим, що ми не вилучали із тексту стоп-слова, власні імена, не проводили лематизацію, не використовували морфологічний аналізатор для того, щоб досліджувати тільки одну певну частину мови, не вилучали слова, що відносяться до загальновживаної лексики.

Висновки. На нашу думку ми отримали задовільний результат. Незважаючи на велику кількість спеціалізованих і міждисциплінарних робіт, присвячених ключовим словами, до теперішнього часу не розроблена послідовна методика виявлення ключових слів людиною. Експериментально підтверджено, що ця операція виконується людьми інтуїтивно. Звідси випливає і складність розробки методів і алгоритмів вилучення ключових слів для обчислювальної техніки. Відсутність чітких формалізованих моделей, надзвичайно розмиті визначення з точки зору комп'ютерної лінгвістики та інших інженерних дисциплін ускладнюють створення і верифікацію відповідного інструментарію.

Список літератури

1. Сысоев П.В. *Лингвистический корпус в методике обучения иностранным языкам*. URL: https://cyberleninka.ru/article/n/lingvisticheskiy-korpus-v-metodike-obucheniya-inostrannym-yazykam (дата звернення: 17.02.2018).
2. *American National Corpus*. URL: https://en.wikipedia.org/wiki/American_National_Corpus (дата звернення: 01.03.2018).
3. *Bank of English*. URL: https://en.wikipedia.org/wiki/Bank_of_English (дата звернення: 01.03.2018).
4. *Британский национальный корпус*. URL: https://ru.wikipedia.org/wiki/Британский_национальный_корпус (дата звернення: 01.03.2018).
5. *Корпус современного американского английского языка*. URL: https://ru.wikipedia.org/wiki/Корпус_современного_американского_английского_языка (дата звернення: 01.03.2018).
6. *International Corpus of English*. URL: https://en.wikipedia.org/wiki/International_Corpus_of_English (дата звернення: 01.03.2018).
7. *Oxford_English_Corpus*. URL: https://en.wikipedia.org/wiki/Oxford_English_Corpus (дата звернення: 01.03.2018).
8. *Национальный корпус шотландского языка*. URL: https://ru.wikipedia.org/wiki/Национальный_корпус_шотландского_языка (дата звернення: 01.03.2018).
9. *Корпусная лингвистика*. URL: http://komiwiki.syktsu.ru/index.php/Корпусная_лингвистика (дата звернення: 05.03.2018).
10. Захаров В.П., Богданова С.Ю. *Корпусная лингвистика*. Иркутск: ИГЛУ, 2011.161 с.
11. *Описательная статистика*. URL: https://ru.wikipedia.org/wiki/Описательная_статистика (дата звернення: 06.03.2018).
12. Николюкин А.Н. *Литературная энциклопедия терминов и понятий*. Интелвак, 2001. 1600 с.
13. Ванюшкин А.С., Гращенко Л.А Методы и алгоритмы извлечения ключевых слов. *Новые информационные технологии в автоматизированных системах*. 2016. С. 85-93.
14. Ahmad K., Gillam L., Tostevin L. University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval. *Proceedings of 8h Text Retrieval Conference (Trec-8)*. 1999. 8 p.

References (transliterated)

1. Sysoev P.V. *Lingvisticheskij korpus v metodike obuchenija inostrannym jazyka* [Linguistic corpus in the methodology of teaching foreign languages]. Available at: https://cyberleninka.ru/article/n/lingvisticheskiy-korpus-v-metodike-obucheniya-inostrannym-yazykam (accessed 17.02.2018).
2. *American National Corpus*. Available at: https://en.wikipedia.org/wiki/American_National_Corpus (accessed 01.03.2018).
3. *Bank of English*. Available at: https://en.wikipedia.org/wiki/Bank_of_English (accessed 01.03.2018).
4. *Britanskij nacional'nyj korpus* [British national corpus]. Available at: https://ru.wikipedia.org/wiki/Британский_национальный_корпус (accessed 01.03.2018).
5. *Korpus sovremennogo amerikanskogo anglijskogo jazyka* [Corpus of Contemporary American English]. Available at: https://ru.wikipedia.org/wiki/Корпус_современного_американского_английского_языка (accessed 01.03.2018).
6. *International Corpus of English*. Available at: https://en.wikipedia.org/wiki/International_Corpus_of_English (accessed 01.03.2018).
7. *Oxford_English_Corpus*. Available at: https://en.wikipedia.org/wiki/Oxford_English_Corpus (accessed 01.03.2018).
8. *Nacional'nyj korpus shotlandskogo jazyka* [Scottish national corpus]. Available at: https://ru.wikipedia.org/wiki/Национальный_корпус_шотландского_языка (accessed 01.03.2018).
9. *Korpusnaja lingvistika* [Corpus linguistics]. Available at: http://komiwiki.syktsu.ru/index.php/Корпусная_лингвистика (accessed 05.03.2018).
10. Zaharov V.P., Bogdanova S.Ju. *Korpusnaja lingvistika* [Corpus linguistics]. Irkutsk: IGLU, 2011.161 p.
11. *Opisatel'naja statistika* [Descriptive statistics] URL: https://ru.wikipedia.org/wiki/Описательная_статистика (accessed 06.03.2018)
12. Nikoljukin A.N. *Literaturnaja jenciklopedija terminov i ponjatij* [Literary encyclopedia of terms and concepts]. Intelvak, 2001. 1600 p.
13. Vanjushkin A.S., Grashhenko L.A Metody i algoritmy izvlechenija kljuchevyh slov [Methods and algorithms for extracting keywords] *Novye informacionnye tehnologii v avtomatizirovannyh sistemah*. 2016. pp. 85-93.
14. Ahmad K., Gillam L., Tostevin L. University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval. *Proc. of 8h Text Retrieval Conf. (Trec-8)*. 1999. 8 p.

Надійшла (received) 23.03.2018

Відомості про авторів / Сведения об авторах / About the Authors

Каніщева Ольга Валеріївна (***Канищева Ольга Валерьевна***, ***Kanishcheva Olga Valeriyivna***) – кандидат технічних наук, доцент, Національний технічний університет «Харківський політехнічний інститут», доцент кафедри інтелектуальних комп'ютерних систем; м. Харків, Україна; ORCID: https://orcid.org/0000-0002-9035-1765; e-mail: kanichshevaolga@gmail.com

Чухненко Маргарита В'ячеславівна (***Чухненко Маргарита Вячеславовна***, ***Chukhnenko Marharyta Vyacheslavivna***) – Національний технічний університет «Харківський політехнічний інститут», магістр; м. Харків, Україна; ORCID: https://orcid.org/0000-0001-5048-7310; e-mail: chukhnenko_margarita@rambler.ru